

# Randomization Resilient To Sensitive Reconstruction

Ke Wang      Chao Han  
 School of Computing Science  
 Simon Fraser University  
 British Columbia, Canada  
 wangk,hanchao@cs.sfu.ca

Ada Waichee Fu  
 Department of Computer Science and  
 Engineering  
 Chinese University of Hong Kong  
 adafu@cse.cuhk.edu.hk

## ABSTRACT

With the randomization approach, sensitive data items of records are randomized to protect privacy of individuals while allowing the distribution information to be reconstructed for data analysis. In this paper, we distinguish between reconstruction that has potential privacy risk, called *micro reconstruction*, and reconstruction that does not, called *aggregate reconstruction*. We show that the former could disclose sensitive information about a target individual, whereas the latter is more useful for data analysis than for privacy breaches. To limit the privacy risk of micro reconstruction, we propose a privacy definition, called  $(\epsilon, \delta)$ -reconstruction-privacy. Intuitively, this privacy notion requires that micro reconstruction has a large error with a large probability. The promise of this approach is that micro reconstruction is more sensitive to the number of independent trials in the randomization process than aggregate reconstruction is; therefore, reducing the number of independent trials helps achieve  $(\epsilon, \delta)$ -reconstruction-privacy while preserving the accuracy of aggregate reconstruction. We present an algorithm based on this idea and evaluate the effectiveness of this approach using real life data sets.

## 1. INTRODUCTION

Randomization is one of the promising approaches in privacy-preserving data mining. With this approach, sensitive data items in records are randomized to protect the privacy of individuals while allowing the distribution information to be reconstructed with reasonable accuracy. An early use of randomization is *randomized response* (RR) for collecting responses on sensitive questions [19]. For example, to find the percentage of employees stealing from the company, the employer asks each employee the question “do you steal from the company?”. To prevent linking the responder to his/her sensitive response, each employee submits the true answer (“Yes” or “No”) with a certain *retention probability*  $p$  and submits

an answer chosen from  $\{Yes, No\}$  at random with probability  $(1 - p)/2$ . This type of randomization, also called *input perturbation*, is extended to categorical values in *privacy preserving data mining* for mining association rules [8, 2, 9, 17]. Randomization is also studied in *privacy preserving data publishing* where a data publisher has collected the original data  $D$  and wants to release a sanitized version  $D^*$  for data mining [3, 11, 16, 22, 4].

In this paper, we consider the data publishing scenario in which the data set  $D$  contains both non-sensitive attributes (e.g., age, gender, etc.) and a sensitive attribute (e.g., disease), as in most realistic settings. We assume that an adversary has named a *target individual*,  $t$ , whose record is contained in  $D$ , and has figured out somehow the non-sensitive attributes of  $t$ . The adversary’s goal is to infer the sensitive attribute of  $t$ . To preserve the privacy of individuals, the sensitive attribute value in each record is randomized following a certain retention probability  $p$ , while allowing reconstruction of distribution information such as the count of records in  $D$  satisfying a given predicate  $\varphi$ . We show that, with the help of non-sensitive attributes, the adversary could reconstruct the distribution of the sensitive attribute for a target individual, even if major privacy definitions are satisfied. If this distribution is skewed, the target individual’s privacy is breached. This attack is termed “reconstruction attack”.

### 1.1 Reconstruction Attacks

One major privacy definition is limiting the change in adversary’s confidence in the sensitive value  $x$  of a given record as a result of interacting with or exposure to the database. For example, the  $\rho_1$ - $\rho_2$  *privacy* proposed in [8] states that if the prior probability  $\Pr[X = x]$  is not more than  $\rho_1$ , the posterior probability  $\Pr[X = x | Y = y]$ , given the published data  $D^*$ , should not be more than  $\rho_2$ , where  $\rho_1 < \rho_2$  and  $X$  and  $Y$  are the variables for the original and perturbed sensitive values in a record, respectively. In the literature [8, 3, 2, 22, 4],  $\Pr[X = x]$  is measured by the fraction of records with  $X = x$  in the *whole* table  $D$ , and  $\Pr[X = x | Y = y]$  is measured by the fraction of records with  $X = x$  among the records with  $Y = y$  in the *whole* table  $D^*$ . Precisely,

$$\Pr[X = x | Y = y] = \frac{\Pr[X = x] \cdot p[x \rightarrow y]}{\sum_x \Pr[X = x] \cdot p[x \rightarrow y]}$$

where  $p[x \rightarrow y]$  is the probability that  $x$  is perturbed to  $y$ , and can be determined by the retention probability  $p$ . Note that these measurements do not take into account the

non-sensitive attributes of records in  $D$  or the acquired non-sensitive information about the target individual  $t$ . The next example shows that with non-sensitive information, the adversary could infer the sensitive information of  $t$  with a probability higher than  $\rho_2$ , even if  $\rho_1$ - $\rho_2$  privacy is ensured.

**EXAMPLE 1 (ATTACKS ON  $\rho_1$ - $\rho_2$  PRIVACY).** Let  $D$  contain  $10 \times k$  records over the sensitive attribute *Disease* and the non-sensitive attributes  $\{\text{Gender}, \text{Age}\}$ , where  $k$  is an integer and *Disease* has the domain  $\{x_1, \dots, x_{10}\}$ . Suppose that  $k$  records in  $D$  have  $\text{Gender} = M$  and  $\text{Age} = 30$ , all of which have the value  $x_1$  for *Disease*. Let  $g$  denote this set of records.  $x_2$ - $x_{10}$  are uniformly distributed among the remaining  $9 \times k$  records in  $D$ . Note, for  $1 \leq i \leq 10$ ,  $\Pr[X = x_i] = 10\%$ , and 0.1-0.5 privacy ensures  $\Pr[X = x_i | Y = y] \leq 50\%$  for all  $x_i$ . This level of privacy can be achieved by retaining the original value in a record with probability 50% and perturbing  $x_i$  randomly to a different value (i.e.,  $\{x_2, \dots, x_{10}\}$ ) with probability  $(1 - 0.5)/9$  [8, 3, 4]. Let  $g^*$  denote the randomized version of  $g$ .

Suppose that an adversary wants to infer the disease of the target individual  $t = \text{Bob}$  having the non-sensitive information  $\text{Gender} = M$  and  $\text{Age} = 30$ . The adversary could estimate the (relative) frequencies of  $x_1, \dots, x_{10}$  in  $g$  based on  $g^*$ , instead of  $D^*$ , because all other records in  $D^*$  do not match  $t$ 's non-sensitive information. Let  $\langle F'_1, \dots, F'_{10} \rangle$  be the estimated frequencies in  $g$ . For a sufficiently large  $g$  (by using a large  $k$ ) and a reasonable estimator such as the maximum likelihood estimator (MLE),  $F'_1$  will be sufficiently close to the true frequency  $f_1$  [2], which is 100%. Consequently, the adversary is able to infer that  $t$  has the disease  $x_1$  with a probability larger than  $\rho_2 = 0.5$ .

A recent breakthrough in privacy definition is *differential privacy* [7]. The idea is hiding the presence or absence of a participant in the database by making two neighbor data sets (nearly) equally probable for giving the produced query answer. Precisely, the  $\lambda$ -differential privacy mechanism ensures that, for any two data sets  $D$  and  $D'$  differing on at most one record, for all queries  $Q$ , and for all query outputs  $o'$ ,

$$\Pr[K(D, Q) = o'] \leq \exp(\lambda) \Pr[K(D', Q) = o']$$

With a small  $\lambda$ ,  $\exp(\lambda)$  is close to 1, so  $D$  and  $D'$  are almost equally likely to be the underlying database that produces the final output of the query. To ensure this property, the  $\lambda$ -differential privacy mechanism adds the noise  $\xi$  to the true answer  $o$  and publishes the noisy answer  $o' = o + \xi$ , where  $\xi$  follows the Laplace distribution  $Lap(b) = \frac{1}{2b} \exp(-\frac{|\xi|}{b})$ ,  $b = 1/\lambda$ . The next example shows that such noisy answers can be exploited to estimate the likelihood of the sensitive value for a target individual.

**EXAMPLE 2 (ATTACKS ON DIFFERENTIAL PRIVACY).** Consider the  $D$  and  $t$  again in Example 1. An adversary could infer the distribution of *Disease* for  $t$  by issuing two queries  $Q_1$  and  $Q_2$ :  $Q_1$  asks for the count of records that satisfy " $\text{Gender} = M \wedge \text{Age} = 30$ " and gets the noisy answer  $o'_1 = o_1 + \xi_1$ , and  $Q_2$  asks for the count of records that satisfy " $\text{Gender} = M \wedge \text{Age} = 30 \wedge \text{Disease} = x_1$ " and gets the noisy answer  $o'_2 = o_2 + \xi_2$ , where  $o_i$  are the true answers and  $\xi_i$  are the noises added,  $i = 1, 2$ . Note that the relative error  $\frac{\xi_i}{o_i}$  gets smaller as the true answer  $o_i$  gets larger, because  $\xi_i$  has the

zero mean and the variance  $2b^2$ , where  $b = 1/\lambda$  is a constant for a given  $\lambda$ -differential privacy mechanism. Therefore, as the answer  $o_1$  increases,  $o'_2/o'_1$  approaches  $o_2/o_1$ , the fraction of records having  $x_1$  among the records that share the gender and age with  $t$ . This discloses the disease  $x_1$  of  $t$  because  $o_2/o_1 = 100\%$ .

In these examples, randomized data or noisy query answers are used to reconstruct the distribution of sensitive information for a target individual, even though strong privacy definitions are satisfied. If such reconstruction is accurate and if the true distribution is skewed, as in these examples, the reconstructed distribution discloses the sensitive information of the target individual with a high probability. This attack is powerful in that it works on different types of randomization techniques and data sharing scenarios, i.e., the random value replacement in Example 1, through either input perturbation or data publishing; random noise addition to query answers in Example 2, also known as *output perturbation*.

## 1.2 Contributions

The contributions in this work are as follows.

- For the first time, we consider the implication of non-sensitive attributes on sensitive reconstruction of data distribution from randomized data. We distinguish two types of reconstruction: The *micro reconstruction* seeks to reconstruct the distribution of the sensitive attribute in a set of records that fully match a target individual on *all* non-sensitive attributes; the *aggregate reconstruction* aims to reconstruct the distribution in a set of records that only partially match a target individual. We argue that micro reconstruction is all we have to be concerned with about privacy risk.
- To address the privacy risk of micro reconstruction, we propose a notion of  $(\epsilon, \delta)$ -*reconstruction-privacy* to ensure a minimum value on the tail probabilities of micro reconstruction error. We present a *bound conversion theorem* that converts between a bound on tail probabilities of a random variable and a bound on tail probabilities of reconstruction error, which allows us to leverage the Chernoff bound to develop a testable instantiation of  $(\epsilon, \delta)$ -reconstruction-privacy. Since the bound conversion theorem does not hinge on the particular form of bounds, our approach can be instantiated to other upper bounds and modified to constrain lower bounds of tail probabilities.
- The promise of this approach is that micro reconstruction is more sensitive to the number of independent trials in the randomization process than aggregate reconstruction, analogous to the fact that the first 10 coin flips are more critical for the estimation of head probability than the second 10 coin flips. We leverage this difference to design an algorithm for achieving  $(\epsilon, \delta)$ -reconstruction-privacy while preserving the utility of aggregate reconstruction.
- Empirical evaluation on real life data sets presents two important findings: Firstly,  $(\epsilon, \delta)$ -reconstruction-privacy is violated even when major privacy definitions such as  $\rho_1$ - $\rho_2$  privacy and differential privacy are satisfied.

Secondly, the additional information loss incurred for achieving  $(\epsilon, \delta)$ -reconstruction-privacy is small.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 defines the problem studied in this work. Section 4 presents an efficient instantiation of  $(\epsilon, \delta)$ -reconstruction-privacy. Section 5 presents the algorithm to achieve  $(\epsilon, \delta)$ -reconstruction-privacy. Section 6 presents empirical findings. Finally, we conclude the paper.

## 2. RELATED WORK

Two classes of randomization methods have been extensively studied in the literature: *random perturbation* and *randomized response*. Random perturbation is primarily used for quantitative data. For example, Agrawal and Srikant [1] build accurate decision tree classification models on the perturbed data, and Kargupta et al. [12] point out that arbitrary randomization can reveal significant amount of information under certain conditions. Randomized response is primarily used for categorical data. Its basic idea was proposed by Warner [19], and based on this technique the problem of mining association rules from disguised data was studied in [8, 9, 17]. In this paper, the term “perturbation” or “randomization” refers to the randomization for categorical data.

Techniques for probabilistic perturbation have also been investigated in the statistics literature. The PRAM method [10] considers the use of Markovian perturbation matrices. The disclosure risk is measured by a notion of *expectation ratios*, defined as the ratio of the expected number of records in the perturbed file with the observed value equal to the value in the original file, and the expected number of records in the perturbed file with the observed value not equal to the value in original file.

Formal definitions of privacy breaches were proposed in [8, 6] following the same paradigm: for every record in the database, the adversary’s confidence in the values of the given record should not significantly increase as a result of interacting with or exposure to the database. Recent works based on such definitions include [16, 11, 3, 2, 22, 4]. These approaches either consider one attribute (i.e., the sensitive attribute) [11], or assume that all attributes are sensitive [16, 3, 2], or ignore the role of non-sensitive attributes in the reconstruction of the sensitive attribute in the context of privacy risk [22, 4]. Reconstruction of data distribution is traditionally considered as utility. To our knowledge, our work is the first to study such reconstruction as privacy breaches.

An alternative to the randomization approach is the partition based approach in which the records are partitioned to ensure some sort of balanced distribution of sensitive data items in each partition [14, 21]. The randomization approach, due to its non-deterministic nature, is more robust to auxiliary information [13, 20, 18].

The differential privacy mechanism [7] hides the presence of a single record in the database by adding random noises to a query answer. As we will see in Section 6, such noise addition is not sufficient to prevent the adversary from reconstructing the distribution of sensitive data for a target individual.

## 3. PROBLEM STATEMENT

We assume that the data publisher has collected a table  $D(NA, SA)$  on *non-sensitive attributes*  $NA = \{A_1, \dots, A_d\}$

and one *sensitive attribute*  $SA$ . Each record in the table corresponds to a participant or individual. For a record  $r$  in  $D$ ,  $r[NA]$  and  $r[SA]$  denote the values of  $r$  on  $NA$  and  $SA$ .  $|\cdot|$  denotes the cardinality of a set. The sensitive attribute  $SA$  has a discrete domain  $\{x_1, \dots, x_m\}$ . The *count* of  $x_i$  refers to the number of records having  $x_i$ , and the *frequency* of  $x_i$  refers to the percentage of records having  $x_i$ . As in [3, 8, 2, 11], we assume that the  $SA$  value in a record is chosen independently at random according to some fixed probability distribution. The publisher allows the researcher to learn this distribution, but wants to hide the  $SA$  value of an individual record.

### 3.1 Perturbation

We consider the data publishing scenario where the data publisher wants to publish  $D$  for data analysis, but wants to hide the  $SA$  value in a record. In the *uniform perturbation* [3, 8, 2, 11], the  $SA$  value  $x$  in a record is processed by flipping a coin with head probability  $0 < p < 1$ , called *retention probability*. If the coin lands on heads,  $x$  is retained; otherwise,  $x$  is replaced with a random value from the domain of  $SA$ , where each value is selected with probability  $(1-p)/m$ . This perturbation process is parameterized by the perturbation matrix  $\mathbb{P}_{m \times m}$ :

$$\mathbb{P}_{ji} = \begin{cases} p + \frac{1-p}{m} & \text{if } j=i \text{ (retain } x_i) \\ \frac{1-p}{m} & \text{if } j \neq i \text{ (perturb } x_i \text{ to } x_j) \end{cases} \quad (1)$$

$p + \frac{1-p}{m}$  is the sum of the probability that  $x_i$  is retained and the probability that  $x_i$  is replaced with the same  $x_i$ . Let  $D^*$  contain all perturbed records. For any subset  $S$  of  $D$ ,  $S^*$  denotes the same set of records as  $S$  in  $D^*$ . Note  $|S^*| = |S|$ . The choice of  $p$  dictates the trade-off between the privacy concern of hiding the sensitive value in a record and the utility for reconstructing the distribution of  $SA$ . The work in [8, 4] determines the maximum retention probability  $p$  for ensuring a given  $\rho_1$ - $\rho_2$  privacy [8] based on  $\rho_1, \rho_2$ , and  $m$ .

The above perturbation process has some interesting properties. First, it modifies only the  $SA$  attribute, not  $NA$  attributes. Therefore, data analysis involving only  $NA$  attributes incurs no information loss by accessing the randomized data  $D^*$ . This is an advantage compared to the differential privacy mechanism [7] where a query answer will be distorted even if it only involves non-sensitive attributes. Second, the perturbation of a record depends on the original  $SA$  attribute in the record, but not on any other records in  $D$ . Therefore, for any subset  $S$  of records from  $D$ , we can assume that  $S^*$  is produced by the same perturbation matrix  $\mathbb{P}$ . This record independence also implies that insertion and deletion of records on  $D$  can be done through insertion and deletion of randomized records on  $D^*$ .

We consider data analysis through answering *count queries*. A count query has a predicate  $\varphi$  of the form  $\wedge(A = a)$ , where  $A$  is either  $SA$  or an attribute in  $NA$ , and  $a$  is a value from the domain of  $A$ . The answer to the query is the count of the records in  $D$  satisfying  $\varphi$ . This answer must be estimated using  $D^*$ . If  $\varphi$  contains no equality for  $SA$ , the answer on  $D^*$  is exactly same as the answer on  $D$ . If  $\varphi$  contains an equality  $SA = x_i$ , a *reconstruction* process will be applied to the subset of records in  $D^*$  that satisfy  $\varphi^-$ , where  $\varphi^-$  is  $\varphi$  with the equality  $SA = x_i$  removed. Let  $S^*$  be this subset and let  $S$  be the set of corresponding records in  $D$ . The reconstruction

seeks the most likely estimator of the distribution of  $SA$  in  $S$ , denoted by  $\overleftarrow{F'}$ , given  $S^*$  and the perturbation operator  $\mathbb{P}$ . The answer for the query is estimated by  $|S|F'_i$ , where  $F'_i$  is the component of  $\overleftarrow{F'}$  for  $x_i$ . The detailed reconstruction will be discussed in Section 4.1.

### 3.2 Adversaries and Micro Reconstruction

We assume that an adversary has named some *target individual*, denoted  $t$ , whose record is contained in  $D$ , and has figured out  $t$ 's values on all non-sensitive attributes  $NA$ . To infer the  $SA$  value of  $t$ , the adversary needs to reconstruct the frequencies  $\overleftarrow{F'}$  of  $SA$  values from the randomized data  $D^*$ . Given the knowledge about  $t$ 's information on all non-sensitive attributes in  $NA$ , the adversary would focus the reconstruction process on the records in  $D^*$  that match all  $t$ 's non-sensitive attributes. The next definition formalizes this reconstruction.

**DEFINITION 1 (MICRO/AGGREGATE RECONSTRUCTION).** A micro group is a set of the records in  $D$  that agree on all attributes in  $NA$ . The micro reconstruction seeks to reconstruct the distribution of  $SA$  in a micro group. For a target individual  $t$ ,  $g_t$  denotes the micro group containing  $t$ 's record and  $g_t^*$  denotes the set of corresponding records in  $D^*$ . An aggregate group is a set of the records in  $D$  that agree on zero or more but not all attributes in  $NA$ . The aggregate reconstruction seeks to reconstruct the distribution of  $SA$  in an aggregate group.

The intent of distinguishing these two types of reconstruction is that micro reconstruction is all we have to be concerned with about privacy risk - aggregate reconstruction does not present privacy risk. The next example illustrates this point.

**EXAMPLE 3.** Let  $NA = \{\text{Gender}, \text{Job}\}$  and  $SA = \text{Disease}$ . Consider a target individual  $t$  (say Bob) with  $\text{Gender} = \text{Male}$  and  $\text{Job} = \text{Teacher}$ . The micro group for  $t$ ,  $g_t$ , contains all records in  $D$  with  $\text{Gender} = \text{Male}$  and  $\text{Job} = \text{Teacher}$ . The micro reconstruction for  $t$  seeks to reconstruct the distribution of  $SA$  in  $g_t$  using the published  $g_t^*$ . This reconstruction is most relevant to  $t$  because  $g_t$  contains all and only the records in  $D$  that match  $t$ 's non-sensitive information. In contrast, aggregate reconstruction involves records that do not match  $t$ 's information in at least one of  $\text{Gender}$  and  $\text{Job}$ , such as (1) all records for  $\text{Job} = \text{Teacher}$ , or (2) all records for  $\text{Gender} = \text{Female}$ , or (3) all records for  $\text{Gender} = \text{Female} \wedge \text{Job} = \text{Teacher}$ , or (4) all records in  $D$ . These reconstructions are less relevant to  $t$  because they are based on more records that do not belong to  $t$ . For example, a high estimated frequency of Breast Cancer in (1) does not mean that  $t$  has a high chance of getting Breast Cancer because most occurrences of Breast Cancer actually come from female teachers.

In the above, we distinguish two types of reconstruction based on the set of records in which the data distribution is estimated. For each type of reconstruction, we can distinguish two types of estimates based on the records used to derive the estimate. In Example 3, we estimate the distribution of  $SA$  in  $g_t$  based on the records in  $g_t^*$ . Alternatively, we can treat  $g_t$  as the difference  $X - Y$  of two sets  $X$  and

$Y$ , where  $S \subseteq X$  and  $Y = X - S$ , and estimate the distribution of  $SA$  in  $g_t$  based on the estimates for  $X$  and  $Y$ . For example, for the set of male teachers,  $g_t$ ,  $g_t = X - Y$ , where  $X$  is the set of all teacher records in  $D$  and  $Y$  is the set of all female teacher records in  $D$ . If  $F'_X$  and  $F'_Y$  are the estimated frequencies of Breast Cancer in  $X$  and  $Y$  based on  $X^*$  and  $Y^*$ , respectively, we can estimate the frequency of Breast Cancer in  $g_t$  by  $(F'_X|X| - F'_Y|Y|)/|g_t|$ . The next definition summarizes these two types of estimation.

**DEFINITION 2 (LOCAL/GLOBAL ESTIMATES).** For any subset  $S$  of  $D$  and any  $SA$  value  $x$ , the local estimate for  $x$  wrt  $S$  is based on the information in  $S^*$ , and a global estimate for  $x$  wrt  $S$  is given by  $(F'_X|X| - F'_Y|Y|)/|S|$ , where  $S \subseteq X$ ,  $Y = X - S$ , and  $F'_X$  and  $F'_Y$  are the local estimates of  $x$  wrt  $X$  and  $Y$ , respectively.

Every local estimate is a global estimate in the special case of  $X = S$  and  $Y = \emptyset$ . At first glance, there is a temptation for considering global estimates because the use of a superset  $X^*$  is in favor of accurate reconstruction. However, we will show that all global estimates are in fact equal to the local estimate in Section 4.1.

Table 1: Notations

Symbols	Meaning
$m$	the domain size $ SA $
$t$	a target individual
$S$	a subset of records in $D$
$S^*$	the corresponding set of $S$ in $D^*$
$g$	a micro group
$g^*$	the corresponding set of $g$ in $D^*$
$x$	a domain value of $SA$
$f$	the frequency of $x$ in $S$
$O^*$	the variable for the observed count of $x$ in $S^*$
$F'$	the variable for the local estimate of $f$
$\overleftarrow{f}, \overleftarrow{F'}, \overleftarrow{O^*}$	the column-vectors of $f, F', O^*$
$\mathbb{P}$	the perturbation matrix in Equation (1)
$p$	the retention probability

### 3.3 Problems

We are now ready to define the problem we will study. We adapt the notation in Table 1 in the rest of the paper. For each target individual  $t$ , the micro reconstruction for  $t$  reconstructs the distribution of  $SA$  most relevant to  $t$ . If the distribution is skewed and if the reconstruction is accurate,  $t$ 's  $SA$  information will be disclosed. To limit this privacy risk, the next definition formalizes a privacy definition through bounding the accuracy of micro reconstruction.

**DEFINITION 3 (( $\epsilon, \delta$ )-RECONSTRUCTION-PRIVACY).** For a micro group  $g$ ,  $g^*$  is ( $\epsilon, \delta$ )-reconstruction-private, where  $\epsilon \geq 0$  and  $\delta \in [0, 1]$ , if for each  $SA$  value  $x$  occurring in  $g$ , whenever  $\Pr \left[ \frac{F' - f}{f} > \epsilon \right] < U$  or  $\Pr \left[ \frac{F' - f}{f} < -\epsilon \right] < L$ ,  $\delta \leq \min\{U, L\}$ , where  $f$  is the frequency of  $x$  in  $g$  and  $F'$  is the variable for a global estimate of  $f$  over the random instances of  $g^*$ .  $D^*$  is ( $\epsilon, \delta$ )-reconstruction-private if  $g^*$  is ( $\epsilon, \delta$ )-reconstruction-private for every micro group  $g$ .

REMARK 1.  $(\varepsilon, \delta)$ -reconstruction-privacy ensures that the (best) upper bounds on tail probabilities for micro reconstruction error greater than  $\varepsilon$  or smaller than  $-\varepsilon$  are not smaller than  $\delta$ . In this sense, the adversary has difficulty to lower the probabilities of a large estimation error. The larger the parameters  $\varepsilon$  and  $\delta$  are, the greater this difficulty is and the more secure the published data is. In this definition,  $\delta$  is a constraint on the upper bounds of tail probabilities (i.e.,  $U$  and  $L$ ). This formulation allows us to leverage the extensive research on upper bounds of tail probabilities in the literature. Alternatively,  $\delta$  could be a constraint on the lower bounds of tail probabilities if such bounds are available, and from Theorem 2, our approach does not hinge on whether  $U$  and  $L$  are upper bounds or lower bounds. In this definition, we consider the estimate  $F'$  estimated from randomized data. In Section 6.3, we will show that the same privacy notion can be applied to  $F'$  estimated from noisy query answers such as those produced by the differential privacy mechanism.

DEFINITION 4 (THE PROBLEM). Given a data set  $D$ , a retention probability  $p$  for randomization,  $\varepsilon$ , and  $\delta$ , where  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , we want to produce a randomized version  $D^*$  that satisfies  $(\varepsilon, \delta)$ -reconstruction-privacy while information for aggregate reconstruction is preserved.

Two main problems are to be solved: how to test if  $(\varepsilon, \delta)$ -reconstruction-privacy is satisfied, and how to achieve  $(\varepsilon, \delta)$ -reconstruction-privacy on a given data set. We answer the first question in Section 4 and answer the second question in Section 5.

## 4. TESTING PRIVACY

We first present an estimation technique for  $F'$  and then present a probabilistic bound for the estimation error of  $F'$ . In the discussion below, the reader is referred to the notations in Table 1.

### 4.1 Maximum Likelihood Estimator

We adapt the *maximum likelihood estimator (MLE)* as our model of local estimates. The next theorem follows from Theorem 2 in [2].

THEOREM 1 (THEOREM 2, [2]). For a subset of records  $S$  and any SA value  $x$ ,  $\overleftarrow{F}'$  computed by  $\mathbb{P}^{-1} \cdot \frac{\overleftarrow{O}^*}{|S|}$  is the maximum likelihood estimator (MLE) of  $\overleftarrow{f}$  in  $S$ , under the constraint  $\Sigma F' = 1$ , where  $\Sigma$  is over all elements of  $\overleftarrow{F}'$ .

In the rest of the paper,  $\overleftarrow{F}'$  denotes the MLE computed by Theorem 1. The presence of the matrix inversion  $\mathbb{P}^{-1}$  makes it troublesome to compute  $\overleftarrow{F}'$  and develop a probabilistic error bound for  $\overleftarrow{F}'$ . The next lemma gives an efficient computation of  $\overleftarrow{F}'$ .

LEMMA 1 (COMPUTING  $\overleftarrow{F}'$ ). For any subset  $S$  of  $D$  and any SA value  $x$ , (i)  $E[O^*] = |S|(fp + (1-p)/m)$ , (ii)  $F' = \frac{O^* / |S| - (1-p)/m}{p}$ , and (iii)  $E[F'] = f$ .

PROOF. (i) Let  $X_k$  be independent and identically distributed (i.i.d.) indicator variables for the event that the  $k$ -th row in  $S^*$  has the SA value  $x$ .  $O^* = \sum_k X_k$ . From

the matrix  $\mathbb{P}$  in Equation (1), if the  $k$ -th row in  $S$  has  $x$ ,  $X_k = 1$  with probability  $p + (1-p)/m$ , and if the  $k$ -th row in  $S$  does not have  $x$ ,  $X_k = 1$  with probability  $(1-p)/m$ . So  $E[O^*] = |S|(p + (1-p)/m) + |S^*|(1-f)(1-p)/m = |S|(fp + (1-p)/m)$ . This shows (i).

(ii) From Theorem 1,  $\overleftarrow{F}' = \mathbb{P}^{-1} \cdot \frac{\overleftarrow{O}^*}{|S|}$ . Let  $[\alpha]_m$  denote a column-vector of the constant  $\alpha$  of the length  $m$ . We have

$$\frac{\overleftarrow{O}^*}{|S|} = \mathbb{P} \cdot \overleftarrow{F}' = p\overleftarrow{F}' + \left[\frac{1-p}{m}\Sigma F'\right]_m = p\overleftarrow{F}' + \left[\frac{1-p}{m}\right]_m$$

The last equation holds because  $\sum F' = 1$  (Theorem 1). Thus,  $\frac{\overleftarrow{O}^*}{|S|} = p\overleftarrow{F}' + \frac{1-p}{m}$ , equivalently,  $F' = \frac{O^* / |S| - (1-p)/m}{p}$ , as required for (ii).

(iii) Taking the mean on both sides of  $F' = \frac{O^* / |S| - (1-p)/m}{p}$ , we get  $E[F'] = \frac{E[O^*] / |S| - (1-p)/m}{p}$ . Substituting  $E[O^*]$  in (i) into the last equation and simplifying, we get  $E[F'] = f$ . This shows (iii).  $\square$

From Lemma 1(ii),  $F'$  can be computed directly from the observed count  $O^*$  without computing the matrix inversion  $\mathbb{P}^{-1}$ . From Lemma 1(iii),  $F'$  is an unbiased estimator of  $f$ . The next lemma shows that, for the MLE model of local estimates, all global estimates are equal to the local estimate.

LEMMA 2. For any subset  $S$  of  $D$  and any SA value  $x$ , every global estimate for  $x$  wrt  $S$  is equal to the MLE for  $x$  wrt  $S$ .

PROOF. From Definition 2, every global estimate wrt  $S$  has the form  $\frac{F'_X |X| - F'_Y |Y|}{|S|}$ , where  $S \subseteq X \subseteq D$  and  $Y = X - S$ , and  $F'_X, F'_Y$  are the MLEs wrt  $S, X, Y$ , respectively. Let  $S^*, X^*, Y^*$  be the sets of records in  $D^*$  corresponding to  $S, X, Y$ , and let  $O^*, O_X^*, O_Y^*$  be the variables for the counts of  $x$  in  $S^*, X^*, Y^*$ , respectively. From Lemma 1(ii),  $F'_X = \frac{O_X^* / |X| - (1-p)/m}{p}$  and  $F'_Y = \frac{O_Y^* / |Y| - (1-p)/m}{p}$ . Substituting these into  $\frac{F'_X |X| - F'_Y |Y|}{|S|}$ , noting  $|S| = |X| - |Y|$  and  $O^* = O_X^* - O_Y^*$ , we get  $\frac{O^* / |S| - (1-p)/m}{p}$ , which is equal to the MLE  $F'$  given by Lemma 1(ii). This shows that every global estimate for  $x$  is equal to the MLE  $F'$  for  $x$ .  $\square$

Consequently, it suffices to consider only local estimates. The next definition refines Definition 3 by considering only local estimates and will be used in the remaining discussion about  $(\varepsilon, \delta)$ -reconstruction-privacy.

DEFINITION 5  $((\varepsilon, \delta)$ -RECONSTRUCTION-PRIVACY (REFINED)). For any micro group  $g$ ,  $g^*$  is  $(\varepsilon, \delta)$ -reconstruction-private, where  $\varepsilon \geq 0$  and  $\delta \in [0, 1]$ , if for each SA value  $x$  occurring in  $g$ , whenever  $\Pr\left[\frac{F' - f}{f} > \varepsilon\right] < U$  or  $\Pr\left[\frac{F' - f}{f} < -\varepsilon\right] < L$ , then  $\delta \leq \min\{U, L\}$ , where  $f$  is the frequency of  $x$  in  $g$  and  $F'$  is the variable for the MLE of  $f$  under the constraint  $\Sigma F' = 1$ .

### 4.2 Probabilistic Error Bounds

A remaining question is how to bound  $\Pr\left[\frac{F' - f}{f} > \varepsilon\right]$  and  $\Pr\left[\frac{F' - f}{f} < -\varepsilon\right]$ . We leverage tail probabilities of random variables in the literature to develop such bounds. Recall that  $O^*$  is the observed count of a SA value and  $F'$  is the reconstructed frequency of a SA value. The next theorem gives a conversion between a probabilistic bound for  $F'$  and a probabilistic bound for  $O^*$ .

**THEOREM 2 (BOUND CONVERSION).** *Consider any subset  $S$  of  $D$  and any SA value  $x$ . Let  $\mu = E[O^*]$ . For any upper tail bound function  $U(\theta, \mu)$  and lower tail bound function  $L(\theta, \mu)$ , and for any comparison operator  $\oplus$  (i.e.,  $<$  or  $>$ ),*

1.  $\Pr \left[ \frac{O^* - \mu}{\mu} > \theta \right] \oplus U(\theta, \mu)$  if and only if  $\Pr \left[ \frac{F' - f}{f} > \varepsilon \right] \oplus U\left(\frac{\varepsilon |S| p f}{\mu}, \mu\right)$ ;
2.  $\Pr \left[ \frac{O^* - \mu}{\mu} < -\theta \right] \oplus L(\theta, \mu)$  if and only if  $\Pr \left[ \frac{F' - f}{f} < -\varepsilon \right] \oplus L\left(\frac{\varepsilon |S| p f}{\mu}, \mu\right)$ .

**PROOF.** We show (1) only because the proof for (2) is similar. From Lemma 1(ii),  $F' = \frac{O^* / |S| - (1-p)/m}{p}$ ,  $O^* = |S|(F'p + (1-p)/m)$ , and from Lemma 1(i),  $\mu = |S|(fp + (1-p)/m)$ . So

$$\begin{aligned} \frac{O^* - \mu}{\mu} > \theta &\Leftrightarrow O^* - \mu > \theta \mu \\ &\Leftrightarrow |S|p(F' - f) > \theta \mu \\ &\Leftrightarrow \frac{F' - f}{f} > \frac{\theta \mu}{|S|p f} = \varepsilon. \end{aligned}$$

These rewriting implies that the probabilities on the two sides of (1) are equal. Then (1) follows because  $\theta = \frac{\varepsilon |S| p f}{\mu}$ .  $\square$

From Theorem 2, if we have a tail probability bound for the error of  $O^*$  (i.e.,  $U(\theta, \mu)$  and  $L(\theta, \mu)$ ), we immediately have a tail probability bound for the error of  $F'$  (i.e.,  $U(\frac{\varepsilon |S| p f}{\mu}, \mu)$  and  $L(\frac{\varepsilon |S| p f}{\mu}, \mu)$ ). Moreover, if the bound for  $O^*$  is the best, the corresponding bound for  $F'$  is also the best (otherwise, a better bound for  $O^*$  can be obtained from Theorem 2). Importantly, the bound conversion does not hinge on the particular form of the bound functions  $U$  and  $L$ . This generality allows us to adapt to the best bounds  $U$  and  $L$  available for  $O^*$  to get the best bounds for  $F'$ .

There is a rich literature on the upper bounds for tail probabilities of random variables. The Markov's inequality applies to any non-negative random variable, therefore, applies to  $O^*$ . The Chebyshev's inequality uses knowledge of the standard deviation to give a tighter bound. However, these bounds are very poor for random variables that fall off exponentially with distance from the mean. The Chernoff bound, due to [5], gives exponential fall-off of probability with distance from the mean. The critical condition that is needed for the Chernoff bound is that the random variable be a sum of independent Poisson trials.

**THEOREM 3 (CHERNOFF BOUNDS, [5, 15]).** *Let  $X_1, \dots, X_n$  be independent Poisson trials such that for  $1 \leq i \leq n$ ,  $X_i \in \{0, 1\}$ ,  $\Pr[X_i = 1] = p_i$ , where  $0 < p_i < 1$ . Let  $X = X_1 + \dots + X_n$  and  $\mu = E[X] = E[X_1] + \dots + E[X_n]$ . For  $\theta \in (0, \infty)$ ,*

$$\Pr \left[ \frac{X - \mu}{\mu} > \theta \right] < U_1(\theta, \mu) = \left( \frac{e^\theta}{(1 + \theta)^{(1 + \theta)}} \right)^\mu \quad (2)$$

and for  $\theta \in (0, 1]$ ,

$$\Pr \left[ \frac{X - \mu}{\mu} < -\theta \right] < L_1(\theta, \mu) = \left( \frac{e^{-\theta}}{(1 - \theta)^{(1 - \theta)}} \right)^\mu \quad (3)$$

These full Chernoff bounds are quite tight but can be clumsy to compute. Using the Taylor series expansion  $\ln(1 + \theta) = \sum_{i \geq 1} (-1)^{i+1} \frac{\theta^i}{i}$  and ignoring higher order terms, the above bounds can be simplified to the following weaker bounds, which covers 95% of cases pretty well: For  $\theta \in (0, \infty)$ ,

$$\Pr \left[ \frac{X - \mu}{\mu} > \theta \right] < U_2(\theta, \mu) = \exp\left(-\frac{\theta^2}{2 + \theta} \mu\right) \quad (4)$$

and for  $\theta \in (0, 1]$ ,

$$\Pr \left[ \frac{X - \mu}{\mu} < -\theta \right] < L_2(\theta, \mu) = \exp\left(-\frac{\theta^2}{2} \mu\right). \quad (5)$$

The Chernoff bound applies to our variable  $O^*$  because  $O^*$  is the sum  $X_1 + \dots + X_n$ , where each  $X_i$  is the indicator variable whether the  $i$ -th row in  $S^*$  has a particular SA value  $x$ , and  $E[O^*] = |S|(fp + (1-p)/m)$  (Lemma 1). Instantiating the upper bounds  $U_i$  and  $L_i$  for  $O^*$  in Equations (2)-(5) into Theorem 2, the next corollary gives the corresponding upper bounds for  $F'$ .

**COROLLARY 1 (UPPER BOUNDS FOR  $F'$ ).** *Let  $U_i$  and  $L_i$  be defined in Equations (2)-(5). For  $\theta \in (0, \infty)$ ,*

$$\Pr \left[ \frac{F' - f}{f} > \varepsilon \right] < U_i(\theta, \mu) \quad (6)$$

and for  $\theta \in (0, 1]$ ,

$$\Pr \left[ \frac{F' - f}{f} < -\varepsilon \right] < L_i(\theta, \mu) \quad (7)$$

where  $\theta = \frac{\varepsilon |S| p f}{\mu}$  and  $\mu = |S|(fp + (1-p)/m)$ .

Corollary 1 gives the concrete upper bounds  $U_i$  and  $L_i$  on the tail probabilities of  $F'$  based on the Chernoff bound. Since these bounds are public,  $(\varepsilon, \delta)$ -reconstruction-privacy implies  $\delta \leq \min\{U_i, L_i\}$ . The question is whether  $\delta \leq \min\{U_i, L_i\}$  is sufficient for  $(\varepsilon, \delta)$ -reconstruction-privacy, in other words, whether there are tighter (i.e., smaller) upper bounds than  $U_i$  and  $L_i$ . To answer this question, we observe from Theorem 2 that any tighter bound for  $F'$  would lead to a tighter bound than the Chernoff bound for  $O^*$ . The fact that the Chernoff bound has been used as the state-of-the-art technique in the past 60 years suggests that it is nontrivial to improve the Chernoff bound. For this reason, we assume that Corollary 1 gives the best upper bounds for  $F'$ ; however, if better bounds on random variables become available, they can be easily adapted through Theorem 2 to obtain better bounds for  $F'$ . This observation leads to the following instantiation of  $(\varepsilon, \delta)$ -reconstruction-privacy based on the Chernoff bound.

**COROLLARY 2 (TESTING  $(\varepsilon, \delta)$ -RECONSTRUCTION-PRIVACY).** *With the upper bounds  $U_i$  and  $L_i$  in Equations (2)-(5), for a micro group  $g$ , for  $\varepsilon \in (0, 1 + \frac{(1-p)/m}{p f}]$  and  $\delta \in [0, 1]$ ,  $g^*$  is  $(\varepsilon, \delta)$ -reconstruction-private if and only if, for every SA value in  $g$ ,*

$$\delta \leq \min\{U_i(\theta, \mu), L_i(\theta, \mu)\} \quad (8)$$

where  $\theta = \frac{\varepsilon |g| p f}{\mu}$  and  $\mu = |g|(fp + (1-p)/m)$ .

The range  $(0, 1 + \frac{(1-p)/m}{p f}]$  of  $\varepsilon$  corresponds to the common range  $(0, 1]$  of  $\theta$  for all of Equations (2)-(5). The condition in Equation (8) can be tested efficiently because all parameters in  $\theta$  and  $\mu$  are known to the data publisher.

## 5. ACHIEVING PRIVACY

We now consider the second major question: how to achieve  $(\varepsilon, \delta)$ -reconstruction-privacy on the published data  $D^*$  for a given data set  $D$ . Corollary 2 gives an efficient condition for  $(\varepsilon, \delta)$ -reconstruction-privacy, but it does not provide a clue on how to achieve this condition if it fails. As the first step towards an answer, we rewrite Equation (8) into a constraint on the size  $|g|$  of a micro group  $g$ . Then we present an algorithm to enforce this constraint. Below, we consider  $(L_1, U_1)$  and  $(L_2, U_2)$  separately.

**THEOREM 4.** *With the upper bounds  $U_1(\theta, \mu)$  and  $L_1(\theta, \mu)$  in Equations (2) and (3), for a micro group  $g$ ,  $\varepsilon \in (0, 1 + \frac{(1-p)/m}{pf}]$ , and  $\delta \in [0, 1]$ ,  $g^*$  is  $(\varepsilon, \delta)$ -reconstruction-private if and only if, for the maximum frequency  $f$  of any SA value occurring in  $g$ ,*

$$|g| \leq \frac{\ln \delta}{w \ln \left( \frac{e^{-\theta}}{(1-\theta)^{(1-\theta)}} \right)} \quad (9)$$

where  $w = fp + (1-p)/m$  and  $\theta = \frac{\varepsilon pf}{w}$ .

**PROOF.** First, we show two claims. Let  $X = \frac{e^\theta}{(1+\theta)^{(1+\theta)}}$  and  $Y = \frac{e^{-\theta}}{(1-\theta)^{(1-\theta)}}$ .

*Claim 1:* for  $\theta \in (0, 1]$ ,  $X \geq Y$ , thus,  $\min\{L_1, U_1\} = L_1$ . Note  $\frac{X}{Y}$  approaches 1 as  $\theta$  approaches 0. To show the claim, it suffices to show that  $\frac{X}{Y}$  is non-decreasing, equivalently, the derivative of  $\frac{X}{Y}$  wrt  $\theta$  is non-negative for  $\theta \in (0, 1]$ . Note

$$\ln \frac{X}{Y} = 2\theta + (1-\theta) \ln(1-\theta) - (1+\theta) \ln(1+\theta)$$

Differentiating both sides wrt  $\theta$ , we get

$$\frac{Y}{X} \left( \frac{X}{Y} \right)' = 2 + [-\ln(1-\theta) + (1-\theta) \frac{-1}{1-\theta}] - [\ln(1+\theta) + (1+\theta) \frac{1}{1+\theta}]$$

and

$$\left( \frac{X}{Y} \right)' = -\frac{X}{Y} \ln(1-\theta^2) \geq 0$$

The last inequality follows because  $X$  and  $Y$  are non-negative and  $\theta$  is in  $(0, 1]$ . This shows Claim 1.

*Claim 2:* for  $\theta \in (0, 1]$ ,  $Y$  is in  $(0, 1)$  and is non-increasing. We show that the derivative of  $Y$  is non-positive (thus,  $Y$  is non-increasing) for  $\theta \in (0, 1]$ . Then the claim follows from the fact that  $Y$  approaches 1 as  $\theta$  approaches 0.

$$\ln Y = -\theta - [(1-\theta) \ln(1-\theta)]$$

Differentiating both sides wrt  $\theta$  gives

$$Y' = Y[-1 - (-\ln(1-\theta) + (1-\theta) \frac{-1}{1-\theta})] = Y \ln(1-\theta) \leq 0$$

The last inequality follows because  $Y$  is non-negative and  $\theta$  is in  $(0, 1]$ . This shows Claim 2.

From Claim 1,  $L_1(\theta, \mu) \leq U_1(\theta, \mu)$ , so Equation (8) degenerates into  $\delta \leq L_1(\theta, \mu) = Y^\mu$ , and  $\ln \delta \leq \mu \ln Y = |g|w \ln Y$ . From Claim 2,  $Y$  is in  $(0, 1)$ , so  $\ln Y < 0$ , and  $|g| \leq \frac{\ln \delta}{w \ln Y}$ . As  $f$  increases,  $w$  and  $\theta = \frac{\varepsilon p}{p + \frac{1-p}{mf}}$  increase, and from Claim 2,  $Y$  is in  $(0, 1)$  and is non-increasing, thus,  $\ln Y$  is decreasing. Since both  $\ln \delta$  and  $\ln Y$  are negative,  $\frac{\ln \delta}{w \ln Y}$  is minimized when  $f$  is maximized. So Equation (8) degenerates into Equation (9).  $\square$

Observe that the right-hand side of the condition in Equation (9) is a constant if the maximum frequency  $f$  is kept unchanged. The idea of our algorithm to enforce this condition is reducing  $|g|$  while keeping  $f$  unchanged. The next theorem gives a similar rewriting based on the bounds  $L_2$  and  $U_2$  in Equations (4) and (5).

**THEOREM 5.** *With the upper bounds  $U_2(\theta, \mu)$  and  $L_2(\theta, \mu)$  in Equations (4) and (5), for a micro group  $g$ ,  $\varepsilon \in (0, 1 + \frac{(1-p)/m}{pf}]$ , and  $\delta \in [0, 1]$ ,  $g^*$  is  $(\varepsilon, \delta)$ -reconstruction-private if and only if, for the maximum frequency  $f$  of any SA value occurring in  $g$ ,*

$$|g| \leq \frac{-2 \ln \delta}{w \theta^2} \quad (10)$$

where  $w = fp + (1-p)/m$  and  $\theta = \frac{\varepsilon pf}{w}$ .

**PROOF.** For  $\theta \geq 0$ ,  $L_2(\theta, \mu) \leq U_2(\theta, \mu)$ , so Equation (8) degenerates into  $\delta \leq L_2(\theta, \mu)$ , where  $\theta = \frac{\varepsilon |g| pf}{\mu}$  and  $\mu = |g|w$ . Note  $\theta = \frac{\varepsilon pf}{w} = \frac{\varepsilon p}{p + \frac{1-p}{mf}}$ . As  $f$  increases,  $\theta$  and  $\mu = |g|w$  increase, hence,  $L_2(\theta, \mu) = \exp(-\frac{\theta^2}{2}\mu)$  decreases. Therefore, it suffices to consider the maximum frequency  $f$  in  $g$  for checking  $\delta \leq L_2$ . The rest of the proof follows from the following rewriting:

$$\delta \leq \exp(-\frac{\theta^2}{2}\mu) \Leftrightarrow \mu \leq -\frac{2 \ln \delta}{\theta^2} \Leftrightarrow |g| \leq \frac{-2 \ln \delta}{w \theta^2}$$

$\square$

In the rest of this section, we develop an algorithm for achieving  $(\varepsilon, \delta)$ -reconstruction-privacy based on Theorem 5, but a similar algorithm can be developed based on Theorem 4. According to Theorem 5, if  $|g| \leq s_g$  fails, where  $s_g = \frac{-2 \ln \delta}{w \theta^2}$ ,  $g^*$  is not  $(\varepsilon, \delta)$ -reconstruction-private. There are several options to restore this inequality. One option is increasing  $s_g$  by reducing either the retention probability  $p$  or the maximum frequency  $f$  in  $g$ . Another option is decreasing  $|g|$  by discarding some records. None of these options is desirable because they either make the data set more random or distort the global data distribution.

Our observation is that  $|g|$  in Equation (10) really refers to the number of independent Poisson trials in the randomization process for generating  $g^*$ . This can be seen from  $\mu = E[O^*] = |g|(fp + (1-p)/m)$  (Lemma 1(i)) where  $|g|$  is the number of indicator variables  $X_k$  for the event that the  $k$ -th row in  $g^*$  has a particular SA value  $x$  (see the proof of Lemma 1). Since the upper bounds in Equations (2)-(5) decrease exponentially in  $\mu$ , reducing  $|g|$  is highly effective to increase these upper bounds, which helps restore the inequality in Equation (10), provided that the frequency  $f$  remains unchanged. At the same time, we want to preserve the frequency of each SA value to minimize the distortion to data distribution. To meet both requirements, we shall randomize a sample  $g_1$  of  $g$  and scale the randomized data  $g_1^*$  back to the original size  $|g|$ . The key is to preserve the frequency of each SA value in both sampling and scaling operations. This task is performed by the following three functions. Assume  $|g| > s_g$ .

1. *Sampling( $g, s_g$ )*: this function takes a sample of the size  $s_g$  from  $g$  such that the number of records for each SA value is reduced by the same fraction. Let  $b = s_g/|g|$

(note  $b < 1$ ). For each  $SA$  value  $x$  occurring in  $g$ , let  $g_1$  contain any  $\lfloor |g_x|b \rfloor$  records from  $g_x$  and one additional record from  $g_x$  with probability  $|g_x|b - \lfloor |g_x|b \rfloor$ , where  $g_x$  denotes the set of records in  $g$  for  $x$ . Note that all records in  $g_x$  are identical. Return  $g_1$ . This step reduces the number of independent trials to  $s_g$  while preserving the frequency of each  $SA$  value.

2. *Perturbing*( $g_1, p, m$ ): this function randomizes the  $SA$  values of the records in  $g_1$  as described in Section 3.1 and returns the randomized  $g_1^*$ .
3. *Scaling*( $g_1^*, |g|$ ): this function scales up  $g_1^*$  to the original size  $|g|$  while preserving the frequency of each  $SA$  value. Let  $b' = |g|/|g_1^*|$ . For each record  $r^*$  in  $g_1^*$ , let  $g_2^*$  contain  $\lfloor b' \rfloor$  duplicates of  $r^*$  and one additional duplicate of  $r^*$  with probability  $b' - \lfloor b' \rfloor$ . Return  $g_2^*$ . Note that the duplication does not increase the number of independent trials because all duplicates of  $t^*$  originate from the same independent trial for  $r^*$ .

The algorithm based on the above idea is described in Algorithm 1. The input consists of  $D, p, m, \epsilon, \delta$  and the output is  $D_2^*$ . For each micro group  $g$ , if  $|g| \leq s_g$ ,  $g_2^*$  is equal to  $g^*$ . Otherwise,  $g_2^*$  is produced by the three steps on Lines 7-9 described above.  $D_2^*$  contains all  $g_2^*$ .

**EXAMPLE 4.** Suppose that a micro group  $g$  contains 5 records for  $x_1$  and 15 records for  $x_2$ .  $|g| = 20$ ,  $|g_{x_1}| = 5$ ,  $|g_{x_2}| = 15$ . Assume  $s_g = 15$ . Since  $|g| > s_g$ , *Sampling*( $g, s_g$ ) produces a sample  $g_1$  of  $g$  as follows.  $b = s_g/|g| = 0.75$ .  $g_1$  contains  $\lfloor 5 \times 0.75 \rfloor = 3$  records from  $g_{x_1}$  and one additional record from  $g_{x_1}$  with probability  $5 \times 0.75 - 3 = 75\%$ ;  $g_1$  contains  $\lfloor 15 \times 0.75 \rfloor = 11$  records from  $g_{x_2}$  and one additional record from  $g_{x_2}$  with probability  $15 \times 0.75 - 11 = 25\%$ . Suppose that after coin flips,  $g_1$  contains 4 records from  $g_{x_1}$  and 11 records from  $g_{x_2}$ . *Perturbing*( $g_1, p, m$ ) produces the randomized version of  $g_1$ ,  $g_1^*$ .

*Scaling*( $g_1^*, |g|$ ) scales up  $g_1^*$  to the size  $|g|$  as follows.  $b' = |g|/|g_1^*| = 20/15 = 1.33$ . For each record  $r^*$  in  $g_1^*$ ,  $g_2^*$  contains  $\lfloor b' \rfloor = 1$  duplicate of  $r^*$  and contains one additional duplicate with probability  $1.33 - 1 = 33\%$ . Suppose that after coin flips, one additional duplicate for  $x_1$  is chosen, and four additional duplicates for  $x_2$  are chosen. So  $g_2^*$  contains 5 records for  $x_1$  and 15 records for  $x_2$ . In general,  $|g_2^*|$  may not be exactly equal to  $|g|$ .

We show that  $D_2^*$  produced by Algorithm 1 satisfies some interesting properties with respect to privacy and utility. Consider a micro group  $g$  such that  $|g| > s_g$ . Let  $g_1, g_1^*, g_2^*$  be computed for  $g$  in Algorithm 1, and let  $O_g^*, O_{g_1}^*, O_{g_2}^*$  be the observed count of a particular  $SA$  value  $x$  in  $g^*, g_1^*, g_2^*$ . Let  $f_g$  and  $f_{g_1}$  be the frequency of  $x$  in  $g$  and  $g_1$ . Let  $F_g', F_{g_1}', F_{g_2}'$  be the MLEs reconstructed from  $g^*, g_1^*, g_2^*$ .  $u \simeq v$  denotes that  $u$  and  $v$  are equal modulo the coin flips in Scaling and Sampling. It is easy to see a few simple facts:

- Fact 1:  $f_{g_1} \simeq f_g$ , that is, Sampling preserves the frequency of  $x$  in  $g$ . This is because the count of every  $x$  in  $g$  is reduced by the same factor  $b$  modulo the coin flips.
- Fact 2:  $O_{g_2}^*/|g_2^*| \simeq O_{g_1}^*/|g_1^*|$ , that is, Scaling preserves the frequency of  $x$  in  $g_1^*$ . This is because each record in  $g_1^*$  is duplicated  $b'$  times modulo the coin flips.

---

#### Algorithm 1 Achieving Reconstruction Privacy

---

Input:  $D, p, m, \epsilon, \delta$

Output: Randomized  $D_2^*$  that is  $(\epsilon, \delta)$ -reconstruction-private

```

1:  $D_2^* \leftarrow \emptyset$ 
2: for all micro groups  $g$  in  $D$  do
3:   compute  $s_g = \frac{-2 \ln \delta}{w \theta^2}$  using Equation (10)
4:   if  $|g| \leq s_g$  then
5:      $g_2^* \leftarrow \text{Perturbing}(g, p, m)$ 
6:   else
7:      $g_1 \leftarrow \text{Sampling}(g, s_g)$ 
8:      $g_1^* \leftarrow \text{Perturbing}(g_1, p, m)$ 
9:      $g_2^* \leftarrow \text{Scaling}(g_1^*, |g|)$ 
10:   add  $g_2^*$  to  $D_2^*$ 
11: return  $D_2^*$ 

```

*Sampling*( $g, s_g$ ):

```

1:  $temp \leftarrow \emptyset$ 
2:  $b \leftarrow s_g/|g|$ 
3: for all  $SA$  value  $x$  occurring in  $g$  do
4:    $g_x \leftarrow$  the set of records in  $g$  having  $x$ 
5:   add to  $temp$  any  $\lfloor |g_x|b \rfloor$  records from  $g_x$ 
6:   add to  $temp$  one additional record from  $g_x$  with probability  $|g_x|b - \lfloor |g_x|b \rfloor$ 
7: return  $temp$ 

```

*Perturbing*( $g_1, p, m$ ):

```

1:  $temp \leftarrow \emptyset$ 
2: for all record  $r$  in  $g_1$  do
3:   let  $r^*$  be  $r$  with  $SA$  perturbed with retention probability  $p$ 
4:   add  $r^*$  to  $temp$ 
5: return  $temp$ 

```

*Scaling*( $g_1^*, |g|$ ):

```

1:  $b' \leftarrow |g|/|g_1^*|$ 
2:  $temp \leftarrow \emptyset$ 
3: for all record  $r^*$  in  $g_1^*$  do
4:   add to  $temp$   $\lfloor b' \rfloor$  duplicates of  $r^*$ 
5:   add to  $temp$  one additional duplicate of  $r^*$  with probability  $b' - \lfloor b' \rfloor$ 
6: return  $temp$ 

```

---

- Fact 3:  $F_{g_1}' \simeq F_{g_2}'$ , that is,  $g_1^*$  and  $g_2^*$  give the same estimate of  $f_g$ . This follows from  $F_{g_i}' = \frac{O_{g_i}^*/|g_i^*| - (1-p)/m}{p}$ ,  $i = 1, 2$  (Lemma 1(ii)) and Fact 2.
- Fact 4:  $E[O_{g_2}^*] \simeq E[O_g^*]$  and  $|g^*| \simeq |g_2^*|$ .  $|g^*| \simeq |g_2^*|$  follows from Sampling and Scaling. From Lemma 1(i),  $E[O_g^*] = |g|(fp + (1-p)/m)$  and  $E[O_{g_1}^*] = s_g(f_1p + (1-p)/m)$ . Since Scaling duplicates each  $x$  occurrence in  $g_1^*$   $\frac{|g|}{s_g}$  times,  $E[O_{g_2}^*] \simeq \frac{|g|}{s_g} E[O_{g_1}^*] = |g|(f_1p + (1-p)/m)$ . Then Fact 1 implies  $E[O_{g_2}^*] \simeq E[O_g^*]$ .

**THEOREM 6 (PRIVACY).** For each micro group  $g$ ,  $g_2^*$  is  $(\epsilon, \delta)$ -reconstruction-private.

**PROOF.** If  $|g| \leq s_g$ ,  $g_2^*$  is  $(\epsilon, \delta)$ -reconstruction-private (Theorem 5). We assume  $|g| > s_g$ .  $g_1^*$  is  $(\epsilon, \delta)$ -reconstruction-private because  $|g_1| \simeq s_{g_1}$  (Theorem 5).  $|g_1| \simeq s_{g_1}$  follows be-



cause  $f_{g_1} \simeq f_g$  (Fact 1) implies  $s_g \simeq s_{g_1}$ , and from  $|g_1| \simeq s_g$ ,  $|g_1| \simeq s_{g_1}$ . Facts 1 and 3 imply  $\frac{F'_{g_2}-f_g}{f_g} \simeq \frac{F'_{g_1}-f_{g_1}}{f_1}$ . So,  $\Pr[\frac{F'_{g_2}-f_g}{f_g} > \varepsilon] \simeq \Pr[\frac{F'_{g_1}-f_{g_1}}{f_1} > \varepsilon]$ , and  $\Pr[\frac{F'_{g_2}-f_g}{f_g} < -\varepsilon] \simeq \Pr[\frac{F'_{g_1}-f_{g_1}}{f_1} < -\varepsilon]$ . Since  $g_1^*$  is  $(\varepsilon, \delta)$ -reconstruction-private, so is  $g_2^*$ .  $\square$

Below, we show that  $F'_2$  has the same mean as  $F'$ . Let  $S$  be any set of micro groups, and let  $S^*$  and  $S_2^*$  be the sets of corresponding records in  $D^*$  and  $D_2^*$ , respectively. For any  $SA$  value  $x$ , let  $F'_2$  denote the estimated frequency of  $x$  in  $S$  based on  $S_2^*$  and let  $F'$  denote the estimated frequency of  $x$  in  $S$  based on  $S^*$ .

**THEOREM 7 (UTILITY).**  $E[F'_2] \simeq E[F']$ .

**PROOF.** Let  $O_2^* = \sum_{g \in S} O_{g_2}^*$  and  $O^* = \sum_{g \in S} O_g^*$ . Let  $|S^*| = \sum_{g \in S} |g^*|$  and  $|S_2^*| = \sum_{g \in S} |g_2^*|$ . From Lemma 1(ii),  $E[F'] = \frac{E[O^*]/|S^*| - (1-p)/m}{p}$  and  $E[F'_2] = \frac{E[O_2^*]/|S_2^*| - (1-p)/m}{p}$ . From Fact 4,  $|S^*| \simeq |S_2^*|$  and  $E[O^*] \simeq E[O_2^*]$ , which implies  $E[F'] \simeq E[F'_2]$ .  $\square$

Despite  $E[F'_2] \simeq E[F']$ ,  $F'_2$  will have a larger error than  $F'$  due to the reduced number of independent trials for  $S_2^*$ . This is exactly what we want in order to restore  $(\varepsilon, \delta)$ -reconstruction-privacy. However, the error increase for aggregate reconstruction is smaller than that for micro reconstruction because aggregation reconstruction involves more than one micro group. We will evaluate this claim empirically in Section 6.

## 6. EMPIRICAL EVALUATION

This empirical study aims to answer two questions: The first question is “to what extent is  $(\varepsilon, \delta)$ -reconstruction-privacy violated assuming that major privacy definitions are satisfied?”. The second question is “what price will be paid for having  $(\varepsilon, \delta)$ -reconstruction-privacy?” Section 6.1 introduces our data sets and utility metrics. Section 6.2 presents the findings in the data publishing setting and Section 6.3 presents the findings in the output perturbation setting.

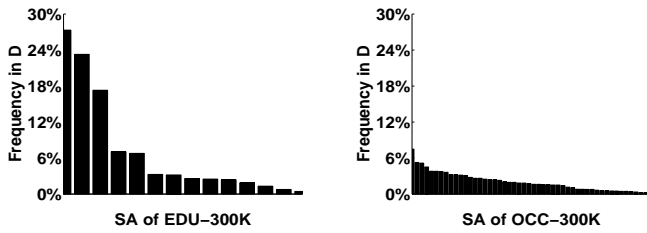


Figure 1: Frequency Distribution for  $SA$

### 6.1 Experimental Setup

**Data Sets.** We utilize the real CENSUS data containing personal information of 500K American adults, previously used in [21],[14], and [4]. Table 2 shows the 7 discrete attributes of the data. Two base tables were generated from CENSUS. OCC denotes the base table with Occupation as the sensitive attribute ( $SA$ ) and the remaining attributes as the non-sensitive attributes ( $NA$ ). EDU denotes

the base table with Education as the sensitive attribute ( $SA$ ) and the remaining attributes as the non-sensitive attributes ( $NA$ ). OCC- $n$  and EDU- $n$  denote the samples of cardinality  $n$ , where  $n = 100K, 200K, 300K, 400K, 500K$ . Figure 1 shows the frequency distribution of  $SA$  for OCC-300K and EDU-300K. EDU-300K has a more skewed distribution than OCC-300K.

Table 2: Number of Values in Attributes

Attributes	Domain Size
Age	77
Gender	2
Education	14
Marital	6
Race	9
Work-class	7
Occupation	50

**Count Queries.** We evaluate the utility of data analysis through count queries of the following form

$$\begin{aligned} & \text{SELECT COUNT } (*) \text{ FROM } D \\ & \text{WHERE } A_1 = a_1 \wedge \dots \wedge A_d = a_d \wedge SA = x_i \end{aligned} \quad (11)$$

where  $\{A_1, \dots, A_d\}$  is a subset of non-sensitive attributes and  $a_j$  is a value from the domain of  $A_j$ ,  $j = 1, \dots, d$ , and  $x_i$  is a value from the domain of  $SA$ . The answer to the query, denoted by  $ans$ , is the count of records in  $D$  that satisfy the predicate in the WHERE clause. Since our primary interest is in aggregate information, we consider only queries that have at least 0.1% selectivity, where the *selectivity* is defined as  $ans/|D|$ . This means that  $d$  is restricted to be 1, 2, or 3 because a query for any larger  $d$  has a selectivity less than 0.1%. We generate a pool of 5,000 queries as follows. For each a query, we randomly select  $d$  from  $\{1, 2, 3\}$  with equal probability and randomly select  $d$  non-sensitive attributes without replacement. For each attribute  $A_i$  selected, we randomly choose a value  $a_i$  from the domain of  $A_i$ . Finally, we randomly choose a value  $x_i$  from the domain of  $SA$  and create a query following the template in Equation (11). If the query has a selectivity of 0.1% or more, we add it to the pool. This process is repeated until the pool contains 5,000 queries.

Table 3: Parameter Table

Parameters	Settings
$p$	0.1, 0.3, <b>0.5</b> , 0.7, 0.9
$\varepsilon$	0.1, 0.3, <b>0.5</b> , 0.7, 0.9
$\delta$	0.14, 0.22, <b>0.3</b> , 0.38, 0.46
$ D $	100K, 200K, <b>300K</b> , 400K, 500K

### 6.2 Findings in Data Publishing

In the data publishing scenario, the randomized data  $D^*$  is published and a query is answered using  $D^*$  and a reconstruction process as described in Section 3.1. Our study focuses on two questions: (i) To what extent is  $(\varepsilon, \delta)$ -reconstruction-privacy violated on  $D^*$ ? (ii) What additional price is incurred for the protection of  $(\varepsilon, \delta)$ -reconstruction-privacy? To answer the first question, we study the *percentage of violating micro groups* in  $D^*$  that fail to satisfy  $(\varepsilon, \delta)$ -reconstruction-privacy. To answer the second question, we measure the (average)

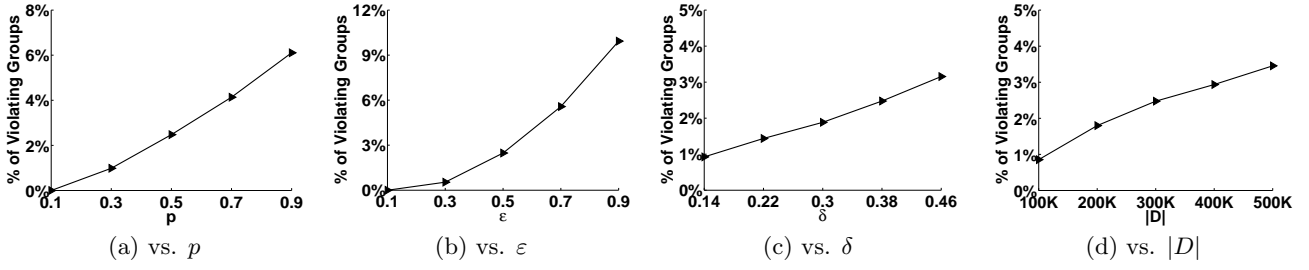


Figure 2: EDU: % of Violating Micro Groups in  $D^*$

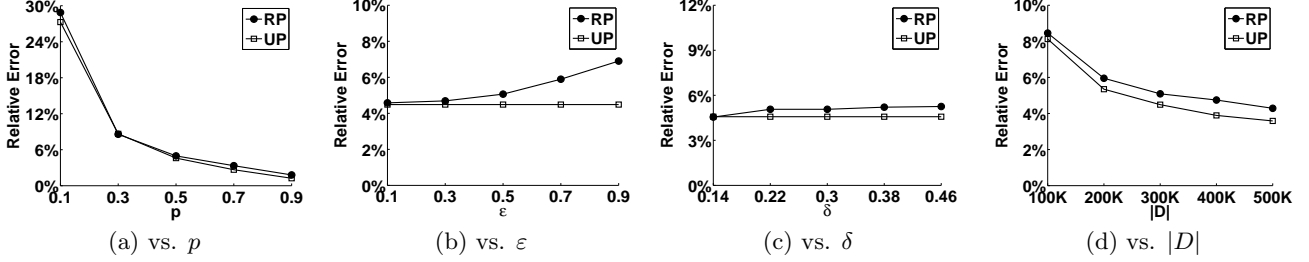


Figure 3: EDU: Comparison of Relative Error for Count Queries

relative error for answering the count queries in our query pool, defined as  $\frac{|est-ans|}{ans}$ , where  $ans$  is the true answer and  $est$  is the estimated answer. We compare the relative error generated using  $D_2^*$  produced by Algorithm 1, denoted by RP (for reconstruction privacy), with the relative error generated using  $D^*$  produced by the standard uniform perturbation, denoted by UP. Both methods use a retention probability  $p$  to randomize the data, thus, ensure some uncertainty of the  $SA$  value in a record such as  $\rho_1$ - $\rho_2$  privacy. According to [8, 3, 4], the maximum  $p$  for providing  $\rho_1$ - $\rho_2$  privacy is  $p = \frac{\gamma-1}{m-1+\gamma}$ , where  $\gamma = \frac{\rho_2}{\rho_1} \times \frac{1-\rho_1}{1-\rho_2}$  and  $m = |SA|$ . Additionally, RP has the parameters  $\varepsilon$  and  $\delta$  for  $(\varepsilon, \delta)$ -reconstruction-privacy. We consider the settings of  $p$ ,  $\varepsilon$ ,  $\delta$ , and  $|D|$ , shown in Table 3. The default settings are in boldface.

### 6.2.1 Findings on EDU Data Sets

Figure 2 shows the percentage of violating micro groups in  $D^*$  vs  $p$ ,  $\varepsilon$ ,  $\delta$ , and  $|D|$ . Here are several observations. Firstly, there is a nontrivial percentage of micro groups that violate  $(\varepsilon, \delta)$ -reconstruction-privacy. A larger retention probability  $p$  leads to more violating micro groups. The violation diminishes when  $p$  becomes very small (i.e., less than 20%), but in this case aggregate reconstruction is affected significantly because  $D^*$  is too noisy, as shown by the larger relative error. A larger  $\varepsilon$  or  $\delta$  leads to more violating micro groups due to a more restrictive privacy constraint. A larger data cardinality  $|D|$  leads to more violating micro groups. This is because a larger  $|D|$  leads to a larger  $|g|$ , i.e., more independent trials when generating  $g^*$ , thus, a more accurate reconstruction. In fact,  $|g| \leq \frac{-2 \ln \delta}{w \theta^2}$  is more likely to be violated as  $|g|$  increases.

For each experiment in Figure 2, Figure 3 shows the relative error of UP and RP. Note that, in Figure 3 (b)(c), UP

remains constant because UP does not depend on  $\varepsilon$  and  $\delta$ . The most significant finding is that, across all of  $p$ ,  $\varepsilon$ ,  $\delta$ , and  $|D|$ , the error of RP is only slightly more than the error of UP. This point can also be seen by cross-examining Figure 2 and Figure 3: the increase of error for RP is much slower than the increase in the percentage of violating micro groups. The reason is that the error boosting of RP through reducing the number of independent trials has less effect on queries that involve a large set of records. This finding supports our claim that the proposed method does not compromise the utility of aggregate information. For  $p$  and  $|D|$ , the trend in Figure 2 and Figure 3 is opposite: as  $p$  or  $|D|$  increases, the percentage of violating micro groups increases, but the error of estimated query answers decreases. This makes sense because violating micro groups are caused by high accuracy of estimated query answers.

### 6.2.2 Findings on OCC Data Sets

We performed a similar study on the more balanced OCC data sets. Figure 4 shows the percentage of violating micro groups and Figure 5 shows the relative error, respectively. As we can see, the findings are quite similar to those of EDU data sets.

## 6.3 Findings on Output Perturbation

Although Definition 3 is based on reconstruction from a randomized data  $D^*$ , the notion of  $(\varepsilon, \delta)$ -reconstruction-privacy is applicable to any reconstruction. In this experiment, we consider reconstruction from noisy query answers in the output perturbation scenario. We assume that differential privacy [7] is in place. The  $\lambda$ -differential privacy mechanism adds random noises  $\xi$  to the query answer  $o$  and publishes the noisy answer  $o' = o + \xi$ , where  $\xi$  follows the Laplace distribu-

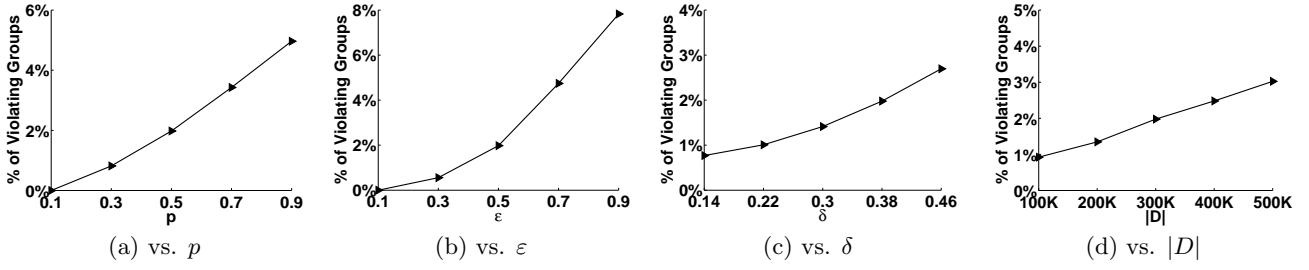


Figure 4: OCC: % of Violating Micro Groups in  $D^*$

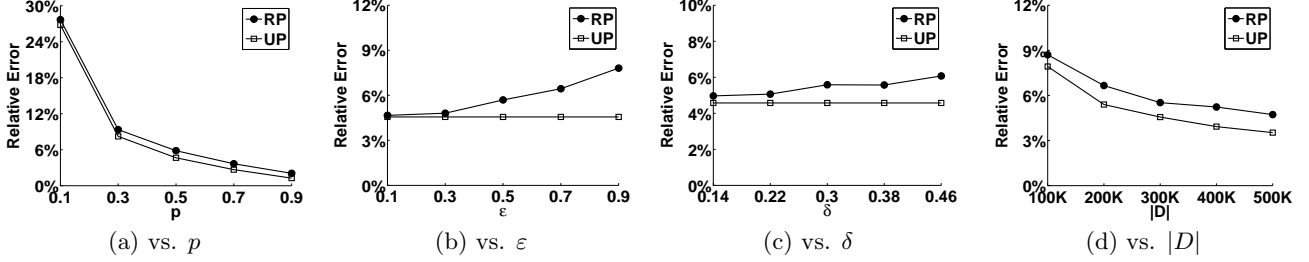


Figure 5: OCC: Comparison of Relative Error for Count Queries

tion  $Lap(b) = \frac{1}{2b} \exp(-\frac{|x|}{b})$ ,  $b = 1/\lambda$ .  $\lambda$  determines the noise level. We show that even if differential privacy is satisfied, there is a concern about violation of  $(\epsilon, \delta)$ -reconstruction-privacy. We use EDU-500K and OCC-500K.

For each data set, we pick 7 micro groups  $g$  that have the largest maximum frequency  $f$  of any  $SA$  value, among those with  $|g| > 70$  for EDU-500K and  $|g| > 100$  for OCC-500K. For each of these groups,  $g$ , let  $f$  and  $F'$  be the true and estimated frequencies of the most frequent  $SA$  value  $x$  in  $g$ .  $F'$  is computed by the noisy answers to two queries  $Q_1$  and  $Q_2$  constructed similar to those in Example 2. Let  $o_i$  be the true answer and let  $o'_i$  be the noisy answer for  $Q_i$ ,  $i = 1, 2$ .  $f = o_2/o_1$  and  $F' = o'_2/o'_1$ . By treating  $\Pr[\frac{F'-f}{f} > \epsilon]$  and  $\Pr[\frac{F'-f}{f} < -\epsilon]$  as the upper bounds of these probabilities themselves,  $(\epsilon, \delta)$ -reconstruction-privacy is violated if  $\Pr[\frac{F'-f}{f} > \epsilon] < \delta$  or  $\Pr[\frac{F'-f}{f} < -\epsilon] < \delta$ . To compute these probabilities, we generated the noisy answers  $o'_1$  and  $o'_2$  100 times and considered the fraction of the cases for  $\frac{F'-f}{f} > \epsilon$  and  $\frac{F'-f}{f} < -\epsilon$ . The numbers for these cases are in Tables 4 and 5.

Take Group 7 in Table 4 (in boldface) for EDU-500K as an example. For  $\lambda = 0.1$  and  $\epsilon = 0.3$ , there are 8 cases for  $> \epsilon$  and 8 cases for  $< -\epsilon$ . Intuitively, this says that, out of the 100 noisy answers ( $o'_1, o'_2$ ) examined, 8 cases have an error greater than 30% and 8 cases have an error less than -30%. In other words, the estimate  $F'$  falls within the  $\pm 30\%$  interval with the confidence level of 84%. The privacy concern comes from the fact that the frequency of  $x$  in  $g$  is more than 70% (shown in the column “ $f$  in  $g$ ”), which is significantly higher than the 2.5% in the whole data set  $D$  (shown in the column “ $f$  in  $D$ ”). Thus, even if the  $\pm 30\%$  interval is large,  $F'$  discloses a much higher probability of having  $x$  for the individuals in  $g$  than

for the individuals in  $D$ . Similar disclosures are observed on the more balanced OCC-500K. For  $\lambda = 0.1$  and  $\epsilon = 0.2$ , Group 2 in Table 5 (in boldface) shows a  $\pm 20\%$  error interval with the confidence level of 84%. Although the frequency  $f$  of  $x$  in this group is only 47%, it is significantly higher than the frequency of 2.4% in the whole data set. Therefore,  $F'$  discloses quite a bit about the  $SA$  value of the individuals in this group.

At  $\lambda = 0.05$ , a larger error for  $F'$  has been observed due to the increased noise level. However, since  $\lambda$  is a constant for a given  $\lambda$ -differential privacy mechanism, the error for  $F'$  can be reduced by a sufficiently large group size  $|g|$  and frequency  $f$  in  $g$ . To provide  $(\epsilon, \delta)$ -reconstruction-privacy, the  $\lambda$ -differential privacy mechanism has to employ a very small  $\lambda$ . This solution shares the same drawback with the solution of using a small retention probability  $p$ , i.e., choosing the global noise parameters, i.e.,  $\lambda$  and  $p$ , according to the *worst case* of any micro group in the data set. As discussed in Section 6.2.1, this type of solutions destroys both micro reconstruction and aggregate reconstruction, making the data useless for *all* queries.

## 7. CONCLUSION

Reconstruction of data distribution is traditionally regarded as utility. In this work, we showed that reconstruction could lead to privacy breaches even if major privacy definitions are satisfied. We formalized a privacy definition to address this risk and presented an enforcement solution. A novelty of this work lies at the distinction between reconstruction that has privacy risk and reconstruction that does not. We leveraged this distinction to meet the dual requirement of privacy and utility. Another novelty is the independence on the particular form of the bounds on tail probabilities. Our privacy

**Table 4: EDU-500K: the Number of Cases for  $\frac{F'-f}{f} > \epsilon$  and  $\frac{F'-f}{f} < -\epsilon$**

Micro Group g	g	f in g	f in D	$\lambda = 0.1$				$\lambda = 0.05$			
				$\epsilon = 0.2$		$\epsilon = 0.3$		$\epsilon = 0.2$		$\epsilon = 0.3$	
				$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$
1	89	0.87	0.025	18	13	14	7	34	29	24	22
2	74	0.77	0.025	25	23	14	7	32	32	28	23
3	138	0.76	0.172	11	9	8	3	18	27	20	15
4	104	0.76	0.172	21	12	9	6	35	28	22	22
5	104	0.75	0.172	23	14	11	6	35	25	21	28
6	77	0.74	0.025	26	11	18	11	26	39	27	26
<b>7</b>	<b>102</b>	<b>0.72</b>	<b>0.025</b>	<b>18</b>	<b>13</b>	<b>8</b>	<b>8</b>	<b>32</b>	<b>31</b>	<b>29</b>	<b>21</b>

**Table 5: OCC-500K: the Number of Cases for  $\frac{F'-f}{f} > \epsilon$  and  $\frac{F'-f}{f} < -\epsilon$**

Micro Group g	g	f in g	f in D	$\lambda = 0.1$				$\lambda = 0.05$			
				$\epsilon = 0.2$		$\epsilon = 0.3$		$\epsilon = 0.2$		$\epsilon = 0.3$	
				$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$	$> \epsilon$	$< -\epsilon$
1	142	0.48	0.038	13	18	11	4	29	29	27	20
<b>2</b>	<b>213</b>	<b>0.47</b>	<b>0.024</b>	<b>8</b>	<b>8</b>	<b>3</b>	<b>0</b>	<b>23</b>	<b>19</b>	<b>15</b>	<b>11</b>
3	111	0.47	0.026	26	18	16	13	40	30	27	29
4	113	0.45	0.024	28	20	17	8	38	30	23	25
5	153	0.45	0.026	18	15	6	6	20	40	26	21
6	237	0.45	0.024	8	9	3	3	17	21	13	12
7	143	0.44	0.038	12	17	12	6	38	34	27	20

definition is a constraint on the upper bounds of tail probabilities and the Chernoff bound in particular. This formulation allows us to leverage the upper bound literature to develop a concrete solution to the problem identified. However, our approach can be instantiated to other upper bounds and modified to constrain the lower bounds of tail probabilities, thanks to the general form of the bound conversion theorem (Theorem 2).

## 8. REFERENCES

- [1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, 2000.
- [2] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving olap. In *SIGMOD*, 2005.
- [3] S. Agrawal and J. Haritsa. A framework for high-accuracy privacy preserving mining. In *ICDE*, 2005.
- [4] R. Chaytor and K. Wang. Small domain randomization: same privacy, more utility. In *VLDB*, 2010.
- [5] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.
- [6] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, 2003.
- [7] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.
- [8] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.
- [9] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *SIGKDD*, 2002.
- [10] J.M. Gouweleeuw, P. Kooiman, L.C.R.J Willenborg, and P.P.de Wolf. Post randomisation for statistical disclosure control: theory and implementation. In *Research paper no. 9731, Statistics Netherlands*, 1997.
- [11] Z. Huang and W. Du. Optrr: optimizing randomized response schemes for privacy-preserving data mining. In *ICDE*, 2008.
- [12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *ICDM*, 2003.
- [13] D. Kifer. Attacks on privacy and definetti’s theorem. In *SIGMOD*, pages 127–138, 2009.
- [14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: privacy beyond k-anonymity. In *ICDE*, 2006.
- [15] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.
- [16] V. Rastogi, S. Hong, and D. Suciu. The boundary between privacy and utility in data publishing. In *VLDB*, 2007.
- [17] S. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, 2002.
- [18] Y. Tao, X. Xiao, J. Li, and D. Zhang. On anti-corruption privacy preserving publication. In *ICDE*, pages 725–734, 2008.
- [19] S. L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. In *The American Statistical Association*, volume 60, 1965.
- [20] R. Wong, A. Fu, K. Wang, Y. Xu, J. Pei, and P. Yu. Probabilistic inference protection on uncertain data. In *ICDM*, 2010.
- [21] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In *VLDB*, 2006.
- [22] X. Xiao, Y. Tao, and M. Chen. Optimal random perturbation at multiple privacy levels. *Proc. VLDB Endow.*, 2:814–825, August 2009.